Data Scientist: Ruslan Shulhan
Collaborator: Luka Anicin

PROBLEM IDENTIFICATION: Sentiment Analysis (CNN)

HYPOTHESIS:

What are the possible ways for corporations (small or big) to measure and understand their image status (representation) among their users (customers or followers), with approximately result in accuracy of 65-89%, by avoiding expensive, biased, and time-consuming in-person surveys, through the method of analyzing the 'language' data (reviews or tweets) by the end of October 2021.

CONTENT:

As most of us are aware, the image of the company plays a profound role in its success. If the company has a damaged reputation in some subdivisions like selling the product or service delivery, it harms clients and investor's trust dramatically. Additionally, the bad reputation correlates with the increasing expense of hiring qualified specialists and retaining relationships with existing clients.

Therefore, monitoring your accountability is vital for most companies. Here, we will be focusing on delivering the product that provides that 'service'. As the main data source, we will be using the API of 'Twitter.com'. It is an American microblogging and social networking service on which users post and interact with messages known as 'tweets'.

The application should be able to gather information about the user's experience regarding the targeting company. So, in the end, it will deliver the overall analysis of how the business stands in today's market based on providing services. The prediction

Data Scientist: Ruslan Shulhan
Collaborator: Luka Anicin

should be in the range of 65-89% accuracy. The deployment should be scheduled for the end of October 2021.

CRITERIA FOR SUCCESS:

The application is fully functional by the end of October 2021. It predicts the company's reputation with an accuracy of 65-89%.

THE SLOPE OF SOLUTION SPACE:

The most focus will be spent on careful data pre-processing. It includes but is not limited to retrieving related to the task data from trusted sources. Then, clean it so we can apply 'special' libraries (tools) to parse and remove noise accordingly. Then, we can gather insights that could be well fitted to our models. If the data is well processed, usually it boosts the accuracy of prediction by a significant amount.

CONSTRAINTS WITHIN SOLUTION SPACE:

Data - gathered data could be compromised (less reliable). In other words, it could consist of information that is not related to the task of this project.

Time - data could be hard to pre-process if there are many steps to make the data clean. The application's speed to predict also slows down dramatically because we work with unstructured data that takes a lot of machine computation. Moreover, the source could be not reachable later on. So, we should re-manage the whole cycle according to the new provider.

Data Scientist: Ruslan Shulhan
Collaborator: Luka Anicin

Limited personal - it means we have limited knowledge and necessary (related) experience toward the working task. So, at some stages, extra learning will be necessary in order to fulfill those gaps.

STAKEHOLDER TO PROVIDE KEY INSIGHT:

In this scenario, the position of stakeholders will be taken by any type of companies that have their business account on 'Twitter.com'. Because 'Twitter.com' is our prior source for retrieving data.

KEY DATA SOURCES:

The data for the training and testing model will be used from the well-known website 'Kaggle.com'. The link to the data is provided below as well.

<[Sentiment140 dataset with 1.6 million tweets | Kaggle](#)>